Cell PRESS
Neuron
Perspective

Theory of Mind: A Neural Prediction Problem

Jorie Koster-Hale[1,*] and Rebecca Saxe[1]
[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
*Correspondence: jorie@mit.edu
http://dx.doi.org/10.1016/j.neuron.2013.08.020

Predictive coding posits that neural systems make forward-looking predictions about incoming information. Neural signals contain information not about the currently perceived stimulus, but about the difference between the observed and the predicted stimulus. We propose to extend the predictive coding framework from high-level sensory processing to the more abstract domain of theory of mind; that is, to inferences about others' goals, thoughts, and personalities. We review evidence that, across brain regions, neural responses to depictions of human behavior, from biological motion to trait descriptions, exhibit a key signature of predictive coding: reduced activity to predictable stimuli. We discuss how future experiments could distinguish predictive coding from alternative explanations of this response profile. This framework may provide an important new window on the neural computations underlying theory of mind.

Introduction

Social life depends on developing an understanding of other people's behavior: why they do the things they do, and what they are likely to do next. Critically, though, the externally observable actions are just observable consequences of an unobservable, internal causal structure: the person's goals and intentions, beliefs and desires, preferences and personality traits. Thus, a cornerstone of the human capacity for social cognition is the ability to reason about these invisible causes. If a person checks her watch, is she uncertain about the time or bored with the conversation? And is she chronically rude or just unusually frazzled? The ability to reason about these questions is sometimes called having a "theory of mind."

Remarkably, theory of mind seems to depend on a distinct and reliable group of brain regions, sometimes called the "mentalizing network" (e.g., Aichhorn et al., 2009; Saxe and Kanwisher, 2003), which includes regions in human superior temporal sulcus (STS), temporo-parietal junction (TPJ), medial precuneus (PC), and medial prefrontal cortex (MPFC). Indeed, the identity of these regions has been known since the very first neuroimaging studies were conducted. By 2000, based on four empirical studies, Frith and Frith concluded that "Studies in which volunteers have to make inferences about the mental states of others activate a number of brain areas, most notable the medial [pre]frontal cortex [(MPFC)] and temporo-parietal junction [(TPJ)]" (Frith and Frith, 2000). Since then, more than 400 studies of these regions have been published. However, although there is widespread agreement on where to look for neural correlates of theory of mind, much less is known about the neural representations and computations that are implemented in these regions. The problem is exacerbated because these brain regions, and functions, may be uniquely human (Saxe, 2006; Santos et al., 2006). Recent evidence suggests that there is no unique homolog of the TPJ or MPFC (Rushworth et al., 2013; Mars et al., 2013), making it even harder to directly investigate the neural responses in these regions.

In the current review, we import a theoretical framework, predictive coding, from other areas of cognitive neuroscience and explore its application to theory of mind. There has recently been increasing interest in the idea of predictive coding as a unifying framework for understanding neural computations across many domains (e.g., Clark, 2013). In this review, we adapt a version of the predictive coding framework that has been developed for mid- and high-level vision. Like vision, theory of mind can be understood as an inverse problem (Baker et al., 2011; Baker et al., 2009); the challenge is to use the observable evidence (in this case human behaviors and states) to infer the invisible causal structure that gave rise to the evidence (the goals, thoughts, and personality of the individual; Seo and Lee, 2012). Also like vision, theory of mind is a complex cognitive process that depends on many different brain regions with likely distinct computational roles (DiCarlo et al., 2012). We suggest that a predictive coding framework can be used both to shed light on existing data about these brain regions, and to suggest productive new lines of research.

First, we briefly review predictive coding, and sketch a model we believe can serve as an integrative framework for the neuroscience of theory of mind. Second, we provide a selective review of existing neuroimaging studies of theory of mind. Across different stimuli and designs, with correspondingly different social information and predictive contexts, we find a classic signature of a predictive error code: reduced neural response to more predictable inputs. Third, we discuss how to distinguish predictive coding from alternative explanations of this response profile, including differences in attention or processing time. Based on recent neuroimaging experiments in visual neuroscience, we suggest strategies for future experiments to test specific predictions of predictive coding. Finally, we discuss the implications of predictive coding for our understanding of the neural basis of theory of mind.

A Predictive Coding Framework

The central idea of "predictive coding" is that (some) neural responses contain information not about the value of a currently perceived stimulus, but about the difference between the stimulus value and the expected value (Fiorillo et al., 2003; Schultz et al., 1997; Schultz, 2010). This general idea is most familiar from studies of "reward prediction error" in

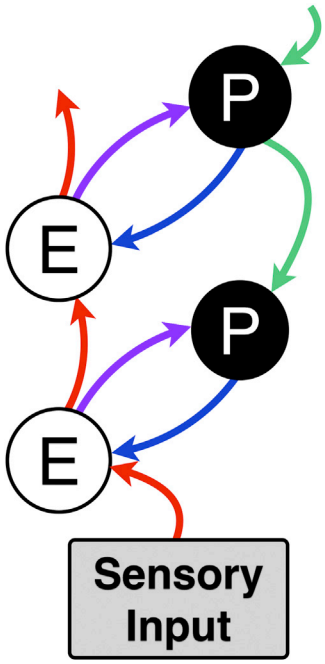

Figure 1. A Sensory-Coding-Based Model of Social Predictive Coding

Predictor neurons (P) code expectations about the identity of incoming input and pass down the prediction to lower level predictor neurons (green arrows) and lower level error neurons (blue arrows). Error neurons (E) act as gated comparators, comparing sensory input from lower levels (red arrows) with the information from predictor neurons (blue arrows). The difference between the predicted input and the actual input is passed up to higher level error neurons, propagating up the processing hierarchy (red arrows). Error neurons also modulate the response of predictor neurons (purple arrows), likely both by inhibiting the predictor neurons making incorrect predictions, and enhancing predictor neurons making correct predictions. When the information that is being passed up from lower levels matches the information carried by the predictor neurons, the error neurons' response to the input is reduced, "explaining away" the predictable input (Rao and Ballard, 1999).

dopaminergic neurons in the striatum. Famously, these neurons initially fire when the animal receives a valued reward, like a drop of juice, and do not respond above baseline to neutral stimuli, such as aural tones. After the animal has learned that a particular tone predicts the arrival of a drop of juice two seconds later, the same neurons fire at the time of the tone. Tellingly, the firing rate of these neurons no longer rises above baseline at the time the juice drop actually arrives. Nevertheless, the neurons still respond to juice. If the tone that typically predicts a single drop of juice is unexpectedly followed by two drops of juice, the neurons will increase their firing; and if the tone is unexpectedly followed by no drops of juice, the neurons *decrease* their firing rate below baseline (Fiorillo et al., 2003; Schultz et al., 1997). These dopaminergic neurons exhibit the simplest and best known example of a neural "error" code: the rate of firing corresponds to any currently "new" (i.e., previously unpredictable) information about the value of coming reward, not to the actual value of any currently perceived stimulus (Bayer and Glimcher, 2005; Nakahara et al., 2004; Tobler et al., 2005).

Predictive coding addresses a general challenge that an animal faces: developing an accurate model of the expected value of all incoming inputs. Thus, predictive coding models can be applied beyond the context of reward prediction to cortical processing more generally. In fact, predictive coding was initially suggested as a model for visual perception (Barlow, 1961; Gregory, 1980; Mumford, 1992), using a visual error code that preferentially encodes unexpected visual information. The key benefit of such a code, proponents suggest, is to increase neural efficiency, by devoting more neural resources to new, unpredictable information.

By contrast to the single population of reward prediction error neurons, predictive coding in the massively hierarchical structure of cortical processing poses a series of challenges. If sensory neurons respond to prediction errors, there must exist other neurons to provide the prediction. Thus predictive coding models require at least two classes of neurons: neurons that formulate predictions for sensory inputs ("predictor" neurons, also called "representation" neurons; Summerfield et al., 2008; Clark, 2013), and neurons that respond to deviations from the predictions ("error" neurons). Because sensory input passes through many hierarchically organized levels of processing (DiCarlo et al., 2012; Felleman and Van Essen, 1991; Logothetis and Sheinberg, 1996; Desimone et al., 1984; Maunsell and Newsome, 1987), a predictive model of sensory processing requires an account of the interactions between prediction and error signals, both within a single level and across levels.

To illustrate the idea, we provide our own sketch of a hierarchical predictive coding model. This proposal is a hybrid of multiple approaches (Friston, 2010; Clark, 2013; Wacongne et al., 2012; de Wit et al., 2010; Spratling, 2010), seems to capture the essential common ideas, and is reasonably consistent with existing data. The key structural idea is that predictor neurons code expectations about the identity of incoming sensory input and pass down the prediction to both lower level predictor neurons and lower level error neurons. Error neurons act like gated comparators: they compare sensory input from lower levels with the information from predictor neurons. When the information that is being passed up from lower levels matches the information carried by the predictor neurons, the error neurons' response to the input is reduced. This type of inhibition is the classic signature of predictive coding, "explaining away" predictable input (Rao and Ballard, 1999). However, when predictor neurons at a higher level fail to predict the input (or lack of input), there is a mismatch between the top-down information from the predictor neurons and the bottom-up information from lower levels, and error neurons respond robustly. This error response propagates up the processing hierarchy. The consequence is a sparse, efficient representation (mostly in predictor neurons) of predictable input, and a robust, distributed response (mostly in error neurons) to unpredictable input, both coordinated across multiple levels of the processing hierarchy (Figure 1).

Within a cortical region, population activity reflects a mixture of responses in the predictor neurons (passing information about predicted inputs down the hierarchy) and the error neurons (passing information about unpredicted inputs up the hierarchy). In principle, predictive coding models need make no assumption about the distribution of these two kinds of neurons within a population; in practice, aggregate population activity is often

dominated by error neurons (Friston, 2009; Wacongne et al., 2012; Egner et al., 2010; Keller et al., 2012; Meyer and Sauerland, 2009). The result is that the classic signature of predictive coding, reduced activity to predictable stimuli, is typically observed when averaging across large samples of neurons within a region (Meyer and Olson, 2011; Egner et al., 2010; de Gardelle et al., 2013). However, (as described in more detail below) signatures of the predictor neurons can also be observed; for example, the predictor neurons would likely show increased response when the input matches their predictions (e.g., de Gardelle et al., 2013).

Following work in sensory processing (e.g., Wacongne et al., 2012), in our proposal both error neurons and predictor neurons convey "representational" information, and both are likely tuned to specific stimuli or stimulus features. Predictor neurons, present at each level of the cortical hierarchy, do not code a "complete" representation of the expected stimulus, but only some features or dimensions of the stimulus, at a relevant level of processing. Each set of predictor neurons can explain only those particular features or dimensions of the input, and correspondingly modulates the response in a highly specific subset of error neurons. Error neurons are similarly distributed throughout the cortex and respond to specific stimulus features (Meyer and Olson, 2011; den Ouden et al., 2012), rather than, for example, a single "error region" signaling the overall amount of error or degree to which the observed stimulus is unpredicted (e.g., Hayden et al., 2011). Thus, for example, in the early visual cortex, predictor neurons code information about the predicted orientation and contrast at a certain point in the visual field, and error neurons signal mismatches between the observed orientation and contrast and the predicted orientation and contrast. In IT cortex, predictor neurons code information about object category; error neurons signal mismatches in predicted and observed object category (den Ouden et al., 2012; Peelen and Kastner, 2011).

One consequence of this model is that, typically, the effects of predictions are limited to relatively few levels of the processing hierarchy. To illustrate, expecting to see John walk in the room would lead to predictions of biological motion, a body, and a face (and possibly to specific predictions within each of these domains), thus reducing error responses in neural populations that respond at this level of abstraction. However, these predictions often cannot be effectively or specifically translated into predictions at the level of early visual receptive fields. Thus, while prediction signals may be passed down the entire cortical hierarchy (Clark, 2013; Rao and Ballard, 1999), in many cases the downstream transformation will make the signal too widespread to be informative. For example, differential responses to predictable complex images have been observed in monkey IT (Meyer and Olson, 2011), and in human ventral temporal cortex (den Ouden et al., 2010; Egner et al., 2010), without corresponding effects in lower visual areas. Only when the environment supports specific, low-level predictions (on the scale of e.g., orientation and contrast at specific points in retinotopy) should error signals be observed at lower levels of processing (Alink et al., 2010; e.g., Murray et al., 2002; Weiner et al., 2010).

When there is no relevant prediction available, error neurons act largely as "feature detection" or "probabilistic belief accumulation" neurons (Drugowitsch and Pouget, 2012). This pattern highlights a key difference between predictive coding models developed for sensory versus reward systems (den Ouden et al., 2012). Reward errors are "signed": the presence of an unexpected reward and the absence of an expected reward are signaled by the same neurons changing their firing rates in opposite directions. By contrast, sensory prediction neurons are likely "unsigned": firing rates increase in the presence of unexplained input.

Finally, our approach contrasts with other recent attempts to integrate social cognitive neuroscience and predictive coding. Because predictive coding is most familiar from the context of reward learning, there has been considerable interest in linking predictive coding to social reward learning (Behrens et al., 2009; Jones et al., 2011; Fehr and Camerer, 2007). Social reward learning can mean either using social stimuli (e.g., smiling faces) as reward, or learning about reward based on observation or consideration of others' experiences (Lin et al., 2012; Zhu et al., 2012; Zaki and Mitchell, 2011; Poore et al., 2012; Jones et al., 2011; Izuma et al., 2008; Chang and Sanfey, 2013; see Dunne and O'Doherty, 2013 for a review). Predictive coding may also be an important mechanism for motor control (i.e., anticipating, and explaining away, the consequences of one's own motor actions). Therefore some authors have linked motor predictions to social predictions via the idea of "mirror neurons," or shared motor representations for one's own and others' actions (Brown and Brüne, 2012; Kilner and Frith, 2008; Patel et al., 2012). The current proposal differs from both of these previous approaches by starting with a hierarchical predictive coding framework developed for cortical visual processing and by focusing on theory of mind, and specifically the attribution of internal states like goals, beliefs, and personality traits.

This proposal is of course too general, and leaves many aspects of the model unspecified (some of which we address below). Nevertheless, the basic features of predictive coding described here provide an integrative framework for many findings in the social cognitive neuroscience of theory of mind.

The Sources of Predictions

The social environment—the actions and reactions of other human beings—can be predicted at a range of temporal scales, from milliseconds (where will she look when the door slams?) to minutes (when she comes back, where will she search for her glasses?) to months (will she provide trustworthy testimony in a court-case?). All of these contexts afford predictions of a person's actions in terms of her internal states, but the sources and timescales of the predictions are different. As we describe in the next three sections, many experiments find that neural responses to predictable actions and internal states are reduced, compared to unpredictable actions and states. This common pattern can provide telling clues about the different types, and sources, of predictions. We find that, while all regions show a higher response to unexpected stimuli, what counts as unexpected varies across regions and experiments, suggesting that, at different levels of processing, neural error responses are sensitive to distinct sources of social prediction.

To help clarify the sources of social prediction, we first review three sources of neural predictions typically manipulated in

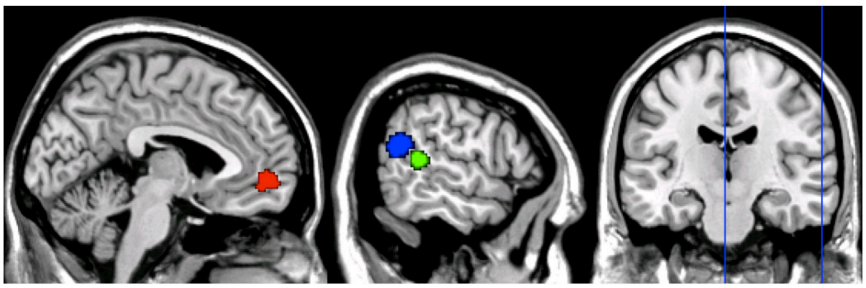

Figure 2. Three Brain Regions Involved in Different Aspects of Theory of Mind: Examples of Individual Regions of Interest

A region in posterior superior temporal sulcus (STS, green, peak voxel [66, −36, 12]) involved in action perception (localized using biomotion relative to scrambled biological motion, Pelphrey et al., 2003); a region in temporo-parietal junction (TPJ, blue, peak voxel [62, −52, 18]) involved in thinking about beliefs and desires (localized using stories about mental states relative to stories about physical events, Saxe and Kanwisher, 2003), and a region in medial prefrontal cortex (MPFC, red, peak voxel [−2, 56, −4]) involved in thinking about people's stable preferences and personalities (localized using attribution of traits to self relative to attribution of traits to someone else, Mitchell et al., 2006). All three ROIs localized using single subject data, $p < 0.001$.

visual cognitive neuroscience experiments. First, given an assumption that the external world is relatively stable, neurons may predict that sensory stimuli will remain similar over short timescales. Predictions based on very recent sensory history can account for increased responses to stimuli that deviate from very recent experience (Wacongne et al., 2012), and reduced responses to stimulus repetition (Summerfield et al., 2008). Predictive coding may therefore offer an account of widespread findings of repetition suppression in neural populations (Grill-Spector et al., 2006). Predictive coding error is consistent with evidence that predictable repetitions elicit more repetition suppression than unpredictable repetitions (Todorovic et al., 2011, Todorovic and de Lange, 2012).

Second, predictable sequences of sensory inputs can be created arbitrarily, through training. For example, Meyer and Olson (2011) created associations between pairs of images; for hundreds of training trials, image A was always presented before image B. After training, the response in IT neurons to image B was significantly reduced when it followed image A. This reduction was highly specific: the response remained high when image B was presented alone, or following some other image, and there was no reduction in the response to image A presented after image B. Other experiments show that a tone can be used as a cue to predict the orientation of an upcoming grating (Kok et al., 2012a), and either a colored frame or an auditory tone can predict whether an upcoming image will be a face or a house (den Ouden et al., 2010; Egner et al., 2010). The reliability of the cue can be stable over the experiment (Egner et al., 2010), or can vary continuously across trials (den Ouden et al., 2010). In all cases, the magnitude of neural responses tracks with the unpredictability of the stimulus, given the cue.

Perhaps most interesting, however, is the third source of predictions: an internal model of the causal structure of the world that generated the observed input (Clark, 2013; Tenenbaum et al., 2011). For example, when two visual bars are presented in alternating positions creating an illusion of motion, the visual system appears to generate an internal model of a single object moving smoothly from one position to the other across the intervening space. As a consequence, the addition of a third bar presented at the right intervening space and time is treated as "predicted," even though that stimulus is otherwise unpredictable within the context of the experiment (Alink et al., 2010).

In principle, all of three these sources of predictions can be applied to social prediction and human actions. In practice, most of the experiments on theory of mind depend on predictions based on prior expectations and an internal model of human behavior (though we do find some evidence of predictions based on temporal proximity). Based on the patterns of findings, we argue that these internal models must be quite abstract, and include expectations that actions will be rational and efficient, and consistent with, for example, the individual's beliefs, personality traits, and social norms.

To reduce the complexity of this literature review, we focus here on three examples of neural responses to actions at three conceptual levels: responses to biological motion and goal-directed action in the superior temporal sulcus (STS), to other people's beliefs and desires in the temporo-parietal junction (TPJ), and to people's stable personality traits in the medial prefrontal cortex (MPFC) (Figure 2). We find that, across all three regions, with respect to the region's preference and level of abstraction, expected stimuli systematically elicit lower activation than unexpected stimuli.

Predicting Goal-Directed Action

The most immediate dimension of the social environment is the visibly observable movements of other people's bodies (e.g., grasping an item, running away) and faces (e.g., gaze shifts, emotional expressions). Brain regions in the superior temporal sulcus (STS) are implicated in many aspects of social action perception, showing robust responses to face and body action in both humans (Jellema et al., 2000; Puce et al., 1998; Bonda et al., 1996; Allison et al., 2002) and monkeys (Perrett et al., 1985; Jellema and Perrett, 2003b; Jellema and Perrett, 2003a). Patients with lesions to the STS have difficulty recognizing actions (Battelli et al., 2003; Pavlova et al., 2003), an effect that is reproduced by creating reversible "lesions" in the STS through repetitive TMS (Grossman et al., 2005). Consistent with a prediction error code, STS response to observed actions is reduced when the observed action can be predicted, and enhanced when the observed action is less predictable. These predictions appear to arise from a variety of sources, ranging from experimental statistics, to constraints on biological motion, to assumption about rational action, suggesting that rather than representing low-level sensory-based statistics, this region represents (and makes predictions about) coherent, rational actions.

First, like many sensory regions, the STS response is sensitive to the recent history of the experiment and is reduced by repetition of a stimulus relevant to human action perception. If two successive images of faces have the same gaze direction (i.e., both gazing right) or the same facial expression (e.g., fearful), the STS response is reduced compared both to a non-repeated presentation and to a repeated presentation of an irrelevant stimulus, such as a house or object (Calder et al., 2007, Ishai et al., 2004, Furl et al., 2007). Similarly, presenting the same action twice in row, from different viewing angles, positions, sizes, and actors leads to reduced STS response relative to a different action (Grossman et al., 2010; Kable and Chatterjee, 2006).

Human action can also be predicted based on internal models at many levels of abstraction, from biomechanics to a principle of rational action. The most basic (and most temporally fine-grained) predictions are constrained by the structure of bones and joints and the forces exerted by muscles. Observers can thus predict the spatiotemporal trajectory of human movements, especially for ballistic motions (Blake and Shiffrar, 2007). Human movements that violate these biomechanical predictions (for example, a finger bending sideways) elicit a higher response than more predictable movements in the STS and related areas (e.g., Costantini et al., 2005). Watching a human-like figure make robot-like, mechanical movements elicits more activity than either a human-like figure making human-like movements or a robot making mechanical movements (Saygin et al., 2012).

Even when they do not violate biomechanical laws, human actions have a typical spatial and temporal structure. Thus, if a person is walking rapidly across the room, we predict that they will continue in the same trajectory, even if they are temporarily occluded. The posterior STS responds more when the person reappears later than expected than when the person emerges at the predicted time; when the person is replaced with a passively gliding object, there is no effect of the time lag (Saxe et al., 2004).

In addition to intrinsic aspects of the action, observers expect others' actions to be temporally and spatially contingent on the structure of the environment. If a bright object flashes near a woman's head, she is very likely to immediately shift her gaze *toward* the object. Seeing the woman immediately shift her gaze *away* from the bright object elicits a higher response in the STS than the predicted gaze shift toward the object (Pelphrey et al., 2003; Pelphrey and Vander Wyk, 2011). This difference is reduced if the woman first waits a few seconds before shifting her gaze, breaking the perception that that flash caused the gaze shift. Similar effects are observed in infants as young as 9 months, using EEG (Senju et al., 2006). In a more extreme mismatch between behavior and environment, watching an agent twisting empty space next to a gear drives a stronger STS response than the agent twisting the gear (Pelphrey et al., 2004).

Finally, the STS internal model of human behavior includes something like a principle of rational action: the expectation that people will tend to choose the most efficient available action to achieve their goal. The same action may therefore be predicted, or unpredicted, depending on the individual's goals and the environmental constraints (Gergely and Csibra, 2003). Correspondingly, the STS response is higher when the same biomechanical action is unpredicted either because it is inefficient, or because it is not a means to achieve the individual's goal.

For example, action efficiency can be manipulated by having a person take a short or long path to the same goal (Csibra and Gergely, 2007), e.g., reaching for a ball efficiently by arching her arm just enough to avoid a barrier, or inefficiently by arching her arm far above the barrier. Across differences in barrier height and arm trajectory, activity in a region of the MTG/STS is correlated with the perceived inefficiency of the action (Jastorff et al., 2011). In a related experiment, observers watch someone performing an unusual action, e.g., a girl pressing an elevator button with her knee. The context renders her action more or less efficient: either her hands are empty, she is carrying a single book, or her arms are completely occupied with a large stack of books. Activity in STS is highest when the action appears least efficient, and lowest when the action appears most efficient (Brass et al., 2007). The STS also responds more to failed actions (e.g., failing to drop a ring onto a peg), an extreme form of inefficiency, than to successful ones (getting the ring onto the peg, Shultz et al., 2011). Predictions for efficient action can even be completely removed from the familiar biomechanics of human body parts: the same inefficient action (going around a non-existent barrier) elicits stronger responses in STS than the efficient version of the same action, when executed by a "worm" (a string of moving dots, Deen and Saxe, 2012).

In other experiments, the STS shows enhanced responses to actions that are unpredicted given the individual's specific goals, even if the action is not inherently inefficient or irrational. For example, an individual who likes (and smiles at) a mug and dislikes (and frowns at) a teddy bear can be predicted to reach for the mug and not the bear. The goal-inconsistent action (reaching for the mug) elicits a higher response in the STS (Vander Wyk et al., 2009). Similarly, when two people are cooperating on a joint action, the STS shows increased responses when one person fails to follow the other's instructions: e.g., when asked to select one specific object (e.g., a red ball), the actor takes the other object (e.g., the white ball; Shibata et al., 2011; see also Bortoletto et al., 2011).

In sum, observers expect human movements to reflect actions, which are sensitive to the environment and efficient means to achieve the individual's goals. These expectations can generate predictions for sequences of movements on the timescale of seconds. All of these sources of predictions can modulate the neural response in the STS, which is reduced when the stimulus fits the prediction.

Predicting Beliefs and Desires

Moving from the scale of seconds to the scale of minutes, the more general version of the principle of rational action is that people will act efficiently to achieve their desires, given their beliefs (Baker et al., 2011). Unlike specific motor intentions, beliefs and desires last from minutes (e.g., the belief that your keys are in your purse) to years (e.g., the desire to become a neurosurgeon). These beliefs and desires can be used to predict aspects of a person's actions, emotions, and other mental

states, especially when the person's beliefs and desires differ from those of the observer (Wellman et al., 2001; Wimmer and Perner, 1983). Among other regions, a brain region posterior to the superior temporal sulcus, in the temporo-parietal junction (TPJ), shows a robust responses while thinking about an individual's beliefs and desires (Saxe and Kanwisher, 2003; Young and Saxe, 2009a; Aichhorn et al., 2009; Perner et al., 2006). If the TPJ includes a prediction error code, it should respond more strongly to beliefs and desires that are unexpected, given the context. Indeed, there is evidence that the TPJ response is reduced when a person's beliefs and desires are predictable (though note that the results reviewed in this section were generally not interpreted in terms of prediction error coding by the original authors). In all of these experiments, the source of prediction is not recent experimental history or trained associations, but rather a high level generative model of human thoughts and behaviors.

One source of predictions about a person's beliefs and desires is their actions (Patel et al., 2012). Observers expect other people to be self-consistent and coherent (e.g., Hamilton and Sherman, 1996). This sensitivity to inconsistencies in belief and action is reflected in the TPJ. For example, in one group of studies, participants read about an act of violent harm or murder. Under the assumption that people usually act in accordance with their beliefs (Malle, 1999), the prediction is that the perpetrator intended the harm; most assaults and murders are not accidental. Next, the participants read about the perpetrator's actual beliefs and desires. Responses in the right TPJ are higher for "unpredicted" innocent or benevolent intentions that exculpate the harm (e.g., she believed the poison was sugar; he only wanted to end the patient's misery from an incurable disease) compared to the "predicted" intention (to kill the person; Buckholtz et al., 2008; Koster-Hale et al., 2013; Yamada et al., 2012; Young and Saxe, 2009b).

Not all actions imply the corresponding intention, however: for example, violation of social norms (e.g., spitting out a friend's cooking back on your plate) are more likely to be committed accidentally than intentionally. Consistent with a prediction error code, the TPJ response is higher for violations of norms performed intentionally ("because you hated the food") versus unintentionally ("because you choked"; Berthoz et al., 2002).

In addition to these general principles, an individual's beliefs and desires can sometimes be predicted based on other information you have about his or her specific group membership and social background. For example, Saxe and Wexler (2005) introduce characters with different social backgrounds, ranging from the mundane (e.g., New Jersey) to the exotic (e.g., a polyamorous cult). Participants then read about that character's beliefs and desires (e.g., a husband who believed it would be either fun or awful if his wife had an affair). The response in right TPJ is reduced for the belief that was predictable, given the character's social background: the person from New Jersey thinking his wife having an affair would be awful, and the person from the polyamorous cult thinking his wife having an affair would be fun. Similarly, when reading about a political partisan, political beliefs that are unexpected, given the individual's affiliation (e.g., a Republican wanting liberal Supreme Court judges) elicits a higher response in right TPJ (Cloutier et al., 2011).

On the other hand, the general plausibility of a belief, in the absence of specific background information about the individual, does not seem to be sufficient to generate a prediction (or a prediction error) in the right TPJ. Without specific background information about the believer, there is no difference in the right TPJ response to absurd versus commonsense beliefs (e.g., "If the eggs are dropped on the table, Will thinks they'll bounce / break," (Young et al., 2010), although the participants themselves rated the absurd beliefs significantly more "unexpected." A possible interpretation of these results is that the internal model of the RTPJ predicts another person's beliefs based on expectations of a coherent individual mind, using information about that individual's specific actions and history, but not based on expectation that others will share one's own beliefs (see below for a contrast with MPFC) or common knowledge. However, since these are null results, they should be interpreted with caution.

In sum, the response in the TPJ to other people's beliefs and desires can be modulated by how predictable those beliefs and desires are, relative to the current environment, the individual's actions, broader social norms, and the individual's specific social background.

Predicting Preferences and Personalities

At even longer timescales, successful prediction of the social environment depends on building distinct models of each of the individual humans who compose one's social group. While some general rules, like the principle of rational action, apply to all people, predicting a specific person's action often depends on knowing the history and traits of that individual. Brain regions on the medial surface of cortex, in both medial prefrontal (MPFC) and medial parietal (PC) cortex, show robust responses while thinking about people's stable personalities and preferences (Mitchell et al., 2006; Schiller et al., 2009; Cloutier et al., 2011). Consistent with a predictive error code, these responses are reduced when new information about a person can be better predicted. Again these predictions appear to be derived from relatively high level expectations that people's traits will be consistent across time and contexts, rather than from local experimental statistics.

Prior knowledge of a person can be acquired through direct interaction. First person experience of another person's traits (e.g., trust-worthiness, reliability), can be manipulated when participants play a series of simple "games" with one or a few other players. By gradually changing the other players' behaviors, it is possible to create parametric "prediction errors." In one experiment, for example, the other player provided "advice" to the participant; this advice shifted over the experiments, so that it was reliable in some phases, and unreliable in others. The response in MPFC tracks with trial-by-trial error in expectations about the informant's reliability (Behrens et al., 2008).

Expectations about other people's traits can also be based on verbal reports and descriptions. For example, the initial behaviors of a (fictional) stranger can create an impression of a certain kind of personality (e.g., "Tolvan gave her brother a compliment"). The MPFC response is enhanced when later actions by the same person are inconsistent with (i.e., unpredicted by) this trait (e.g., "Tolvan gave her sister a slap") compared to

when they are predictable (e.g., "Tolvan gave her sister a hug"; Ma et al., 2012; Mende-Siedlecki et al., 2012).

When specific information about a person's reputation or traits is unavailable, we may predict others' preferences by assuming that they will share our own preferences (Krueger and Clement, 1994; Ross et al., 1977). In one series of studies (Tamir and Mitchell, 2010), participants judged the likely preferences of strangers (e.g., is this person likely to "fear speaking in public" or "enjoy winter sports"?) about whom they had almost no background information. Under those circumstances, the response of the MPFC was predicted by the discrepancy between the attributions to the target and the participant's own preference for the same items: the more another person was perceived as different from the self, for a specific item, the larger the response in MPFC.

In all, human observers appear to formulate predictions for other people's movements, actions, beliefs, preferences, and behaviors, based on relatively abstract internal models of people's bodies, minds, and personalities. These predictions are reflected in multiple brain regions, including STS, TPJ, and MPFC, where responses to more predictable inputs are reduced, and to less predictable inputs are enhanced.

Consistent with our general proposal for prediction error coding, reduced responses to predicted stimuli in these experiments are typically restricted to relatively few brain regions, and by implication, to relatively few levels of the processing hierarchy. Beliefs or actions that are unpredicted, based on high level expectations, do not elicit enhanced responses at every level of stimulus processing (e.g., early visual cortex, word form areas, etc). Nor are prediction errors signaled by a single centralized domain general "error detector." Instead, relatively domain- and content-specific predictions appear to influence just the error response at the relevant level of abstraction.

Beyond Error

In sum, human thoughts and actions can be rendered unexpected in many ways, and across many such variations a common pattern emerges: brain regions that respond to these stimuli also show enhanced responses to "unexpected" inputs. This profile is the classic signature of error neurons, and therefore consistent with a predictive coding model of action understanding.

While *consistent* with predictive coding, however, these results provide only weak evidence in favor of predictive coding. Increased responses to unexpected stimuli can be explained by many different mechanisms, including increased "effort" required, increased attention, or longer evidence accumulation under uncertainty. The predictive coding framework will therefore be most useful if it can make more specific predictions and suggest new experiments.

(1) Distinguishing Prediction from Attention

A salient alternative explanation for enhanced responses to unpredicted stimuli relies on attention. Unexpected stimuli may garner more attention, and increased attention can lead to more processing and higher activation (e.g., Bradley et al., 2003; Lane et al., 1999). Similarly, increased processing effort or longer processing time can predict higher activation (e.g., Cohen et al., 1997). Thus, higher activation to unexpected stimuli could reflect greater attention or longer processing, rather than prediction coding errors. However, relative to these accounts, predictive coding has a distinctive signature.

By hypothesis, predictions codes are more precise, more computationally efficient, and less noisy than error codes (Friston, 2005; Jehee and Ballard, 2009; Rao and Ballard, 1999; Spratling, 2008). As a result, in a predictive coding model, better speed and accuracy of perception are associated with *reduced* overall neural responses to predicted stimuli (Kok et al., 2012a; den Ouden et al., 2009). By contrast, attention may cause better speed and accuracy of performance by *increasing* overall neural responses to attended stimuli (Feldman and Friston, 2010; Friston, 2010; Herrmann et al., 2010; Hillyard et al., 1998; Kok et al., 2012b; Martinez-Trujillo and Treue, 2004; Reynolds and Heeger, 2009; Treue and Martínez-Trujillo, 1999). That is, whereas attention may increase gain in neural responses to the attended stimulus, predictions improve perception by decreasing noise (or increasing sparseness) in neural responses to the predicted stimulus.

If the neural responses described in the previous section reflect prediction error, reduced neural responses should be accompanied by improvements in behavioral performance: people should make judgments more quickly, with less error, and with more sensitivity to expected stimuli. Indeed, behavioral evidence suggests that observers make faster and more accurate judgments about people who behave as expected in social contexts. After watching two people engage in part of a cooperative action or conversation, participants are faster and more accurate when both agents are behaving as expected (e.g., responding aggressively or cooperatively, responding communicatively or non-communicatively, or right away, instead of too early or late; Manera et al., 2011; Neri et al., 2006; Graf et al., 2007). Important next questions will be to look for these signatures in other aspects of social cognition, such as goal inference or belief attribution.

An interesting extension of this idea is the proposal that the sparser prediction signal should also be easier to decode from a neural population than the more distributed error signal, within a single region and task (Kok et al., 2012a; Sapountzis et al., 2010). In an elegant study, Kok et al., (2012a) asked participants to make fine perceptual discriminations between oriented gratings. They hypothesized that when the orientation of the gratings was accurately predicted by a cue, the representation of the grating would be largely in the sparser predictor neurons, whereas when the orientation was not accurately predicted (i.e., on the relatively rare invalidly cued trials), then the representation of the orientation would be largely in the more distributed error neurons. Three predictions of their model were confirmed in the responses of early visual cortex. First, the overall response to the gratings was lower when the orientation was predicted than when it was unpredicted (the classic pattern of "explaining away" the error signal). Second, behavioral discriminations on the gratings were more accurate when the orientation was predicted than when it was unpredicted, consistent with the hypothesis of a more efficient code. Third, and critically, the orientation of the gratings could be more easily decoded from the spatial pattern of neural responses in early visual cortex when the orientation was predicted than when it was

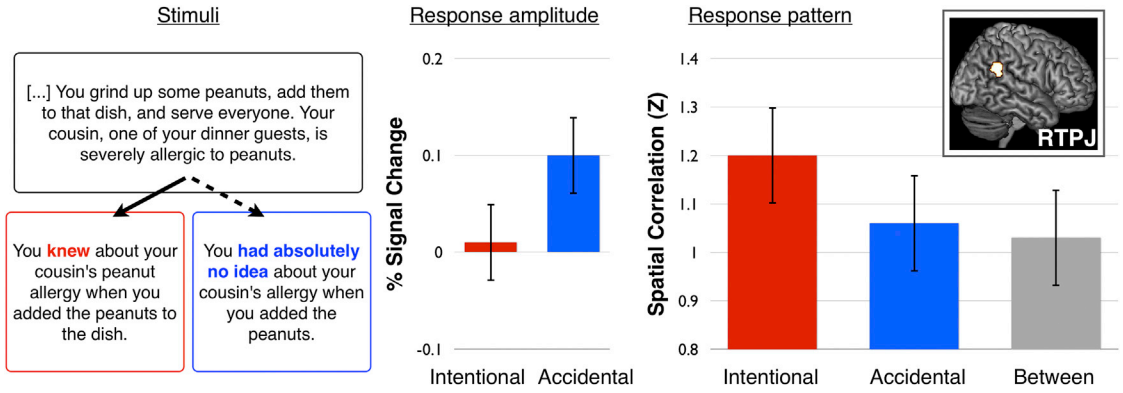

Figure 3. Predicted Stimuli Elicit Reduced Activity with Sharpened Representations

Kok et al. (2012a) report that an accurate prediction led to reduced response amplitude in early visual cortex, but also simultaneously to an "improved" stimulus representation, as measured by multi-voxel pattern analysis. Consistent with this suggestion, we find that in the right temporo-parietal junction (RTPJ; right box) the response amplitude to a predicted belief was lower, but the spatial pattern associated with that belief category was more reliable. Left: a sample stimulus. All stories described first a harmful action, and then the agent's belief. The "predicted" belief (solid arrow) was consistent with the action (i.e., making the act an intentional harm). The "unpredicted" belief (dotted arrow) was inconsistent and rendered the harm an accident. Middle: The amplitude of response in the TPJ was lower for the intentional than accidental condition. Right: The spatial pattern of response in the TPJ was most robust and reliable across trials for intentional harms, and somewhat less reliable for accidental harms. Data from Koster-Hale et al. (2013).

unpredicted, consistent with the hypothesis that the prediction signal is spatially sparser than the error signal. Finally, Kok et al., (2012a) distinguished the effects of prediction from effects of attention, by manipulating the participants' task. Directing attention to the gratings' orientation (versus contrast) improved decoding of orientation in V1, but the effects of attending to orientation, and of seeing the unpredicted orientation, were independent and additive.

A corresponding hypothesis should be easy to test with respect to the neural representation of human behaviors, thoughts and personalities. The lower responses to expected stimuli should be accompanied by better decoding of relevant stimulus dimensions. Indeed, our own results from the TPJ are consistent with this hypothesis. As described above, when reading about harmful actions (e.g., putting poison powder in someone's coffee), the TPJ response is higher to "unpredicted" innocent beliefs (e.g., that the powder was sugar) than to "predicted" consistent beliefs (e.g., that the powder was poison; Young and Saxe, 2009b). We also found that using spatial pattern analysis in the TPJ, we could decode the difference between innocent and guilty beliefs (Koster-Hale et al., 2013). Based on Kok et al., (2012a), a further prediction is that the decoding should be driven by a sparser and more efficient response to the predicted category; and indeed, re-analysis of our data suggests that the guilty beliefs elicit a more distinctive (i.e., more correlated across trials) spatial pattern than the "unpredicted" innocent beliefs (Figure 3).

Interestingly, the benefits of an accurate prediction may be quite specific to the aspects of the stimulus that are accurately predicted. As we suggest earlier, most predictions are limited to a particular level of abstraction; given a high-level prediction, the probability of lower-level features appearing will be too widely distributed to be informative. As a result, accurate predictions may improve behavioral performance (and neural decoding) at the representational level of the prediction (e.g., which object a person wanted) but fail to improve, or even degrade, these measures for lower-level features (e.g., where in space someone looked; He et al., 2012).

(2) Finding the Prediction Signal

An important direction for future research will be to focus on signatures of the predictor neurons, in addition to the error neurons. At least four different strategies may help to identify prediction signals, and distinguish them from the often more dominant error signals.

First, unlike error neurons, predictor neurons should show robust activity when the stimulus fits prior predictions. Consistent with this suggestion, a recent study found that although the majority of voxels in the fusiform face area (FFA, Kanwisher et al., 1997; Kanwisher, 2010) was suppressed for a repeated face, a subset of voxels reliably showed the reverse pattern (de Gardelle et al., 2013), termed repetition enhancement (see also Turk-Browne et al., 2006; Müller et al., 2013). Intriguingly, these two populations of voxels also showed different patterns of functional connectivity. It will be intriguing to test whether the STS, TPJ, PC, or MPFC similarly contain subsets of voxels with enhanced responses to predicted actions or beliefs, and whether these voxels have distinctive patterns of functional connectivity with other regions, especially because unlike face processing, the direction of information flow among regions involved in theory of mind is largely unknown.

Second, because both predictor neurons and error neurons may have preferred stimuli (or stimulus features), it may be possible to identify the content of the prediction independent from the response to the subsequent stimulus. For example, the response of the FFA seems to increase when a face stimulus is predicted, as well as (and partially independent from) when a face stimulus is observed (den Ouden et al., 2010; Egner et al., 2010). Note though that neither of the existing studies could fully independently identify the response to predicting a face, because in both cases, the probability of a face was exactly reciprocal to the probability of the only other possible stimulus, a house. By including a third category of stimulus, or a third

possible cue, or by independently varying the predictive value of the two cues, it should be possible to independently measure category-specific responses to the prediction of a category, versus the response to that category when observed.

Third, and relatedly, predictor neurons can signal the expectation of a stimulus that never occurs. In some cases, the absence of an expected stimulus should generate error activity (den Ouden et al., 2010; Todorovic et al., 2011; Wacongne et al., 2012). For example, the activity pattern in IT generated by the surprising absence of an object contains information about the identity of the absent stimulus (Peelen and Kastner, 2011). Unlike the "signed" (i.e., below baseline) error response in reward systems, sensory neurons thus seem to show an increased response to an unexpectedly absent stimulus (though note that there is some disagreement as to whether this activity is driven only by the prediction signal before the stimulus is expected to appear, or by a combination of the prediction signal with a subsequent error signal when the stimulus fails to appear, e.g., den Ouden et al., 2010).

Fourth, the prediction and the error signals could be separable in time. Specifically, under some circumstances, prediction signals can begin before the stimulus, whereas error signals are typically triggered by the stimulus itself (e.g., Hesselmann et al., 2010). The low temporal resolution of fMRI may make it hard to test this hypothesis directly. However one pattern of results is consistent with the idea that the STS contains predictions of upcoming biological motion: still photographs of a person in mid-motion (such as a discus thrower in the middle of throwing a disc) elicited more activity in the STS than images that do not imply or predict motion (the same discus thrower at rest; Kourtzi and Kanwisher, 2000; Senior et al., 2000).

Fifth, error responses in a single region may be influenced by predictions from different sources, and these different sources may be spatially separable. For example, FFA shows repetition suppression for both repetition of one identical face image (plausibly a very low-level prediction) and for repetition of a face across different sizes (requiring a higher-level prediction). These error signals were related to different patterns of functional connectivity between FFA and lower level regions (Ewbank et al., 2013). By analogy, there may be different patterns of functional correlations related to different sources of prediction for human actions. In one experiment, for example, the STS response was enhanced for actions that were unpredicted for two different reasons: reaching for empty space next to a target (which is an inefficient or failed action), or reaching for a previously nonpreferred object (which is unpredicted relative to an inferred goal; Carter et al., 2011; see also Bubic et al., 2009). It would be interesting to test whether these two kinds of errors are associated with spatially distinct sources of functional connectivity to the STS.

(3) Using Predictive Coding to Test the Neural Computations Underlying Theory of Mind

The framework of predictive coding offers a new opportunity to study the neural representations of others' actions and thoughts, using new experimental designs. The necessary logic has been developed in repetition suppression experiments (Grill-Spector et al., 2006). Complex stimuli elicit responses in many different brain regions simultaneously, making it hard to dissociate the representational and computational contributions of different brain regions. Consequently, in higher level vision, repetition suppression has been used to differentiate the stimulus dimensions or features represented in multiple co-activated regions. For example, although both the FFA and the STS face area show repetition suppression when the identity of a face is repeated, only a more anterior STS region shows a reduced response when the emotional expression is repeated across different faces (Winston et al., 2004).

Looking for prediction error offers a generalized, and more flexible, version of repetition suppression studies; critically, it only requires that a stimulus be surprising along some dimension, without having to repeat the stimulus. This flexibility is particularly advantageous for studies using naturalistic or social stimuli, which can be hard to repeat without also invoking repetition of a number of confounding, but unrelated representations—for example, words, syntactic structure, or faces. Exact repetitions of complex stimuli can be unnatural or pragmatically odd, which may especially limit the ability to study repetition suppression in young or special populations. By contrast, the distribution of observed error signals could reveal both which neural populations or regions are coding the relevant dimensions and features, and what the sources of predictions are.

Finally, and perhaps most importantly, this framework may enrich theorizing about neuroimaging results in social cognitive neuroscience. One of the key challenges facing social cognitive neuroscience is that the richness of the data often surpasses the precision of the theories. This proves to be a problem both for interpreting the data—inverse inferences are very rarely well-constrained enough to be compelling, despite their role in theory building—and for designing new hypotheses and experiments. Increased response in a brain region has been argued to indicate both that the stimulus carries many relevant features to a region and that the stimulus was harder to process or a less good "fit" to the region; this problem is exacerbated when trying to interpret different neural patterns across groups (i.e., special populations). If we can begin to break down (a) what kinds of predictions a region makes, (b) what kind of information directs those predictions, and (c) what constitutes an error, it may be possible to formulate much more specific hypotheses about the computations, and information flow, that underlie human theory of mind.

In sum, we find a predictive coding approach to theory of mind promising. There is extensive evidence of a key signature of predictive coding, in fMRI studies of theory of mind: reduced responses to expected stimuli. Existing data also provide hints of other, more distinctive signatures of predictive coding. Future experiments designed to more directly test the predictions and errors represented in different brain regions may provide an important new window on the neural computations underlying theory of mind.

ACKNOWLEDGMENTS

The authors thank Amy Skerry, Hilary Richardson, Todd Thompson, and Nancy Kanwisher for comments and discussion. The authors gratefully acknowledge support of this project by an NSF Graduate Research Fellowship (#0645960 to JKH) and an NSF CAREER award (#095518), NIH (1R01 MH096914-01A1), and the Packard Foundation (to RS).

REFERENCES

Aichhorn, M., Perner, J., Weiss, B., Kronbichler, M., Staffen, W., and Ladurner, G. (2009). Temporo-parietal junction activity in theory-of-mind tasks: falseness, beliefs, or attention. J. Cogn. Neurosci. *21*, 1179–1192.

Alink, A., Schwiedrzik, C.M., Kohler, A., Singer, W., and Muckli, L. (2010). Stimulus predictability reduces responses in primary visual cortex. J. Neurosci. *30*, 2960–2966.

Allison, T., Puce, A., and McCarthy, G. (2002). Category-sensitive excitatory and inhibitory processes in human extrastriate cortex. J. Neurophysiol. *88*, 2864–2868.

Baker, C.L., Saxe, R., and Tenenbaum, J.B. (2009). Action understanding as inverse planning. Cognition *113*, 329–349.

Baker, C.L., Saxe, R.R., and Tenenbaum, J.B. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In Proceedings of the thirty-second annual conference of the cognitive science society (pp. 2469–2474).

Barlow, H.B. (1961). Possible principles underlying the transformation of sensory messages. In Sensory Communication, W.A. Rosenblith, ed. (Cambridge, MA: MIT Press), pp. 217–234.

Battelli, L., Cavanagh, P., and Thornton, I.M. (2003). Perception of biological motion in parietal patients. Neuropsychologia *41*, 1808–1816.

Bayer, H.M., and Glimcher, P.W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron *47*, 129–141.

Behrens, T.E., Hunt, L.T., Woolrich, M.W., and Rushworth, M.F. (2008). Associative learning of social value. Nature *456*, 245–249.

Behrens, T.E., Hunt, L.T., and Rushworth, M.F. (2009). The computation of social behavior. Science *324*, 1160–1164.

Berthoz, S., Armony, J.L., Blair, R.J.R., and Dolan, R.J. (2002). An fMRI study of intentional and unintentional (embarrassing) violations of social norms. Brain *125*, 1696–1708.

Blake, R., and Shiffrar, M. (2007). Perception of human motion. Annu. Rev. Psychol. *58*, 47–73.

Bonda, E., Petrides, M., Ostry, D., and Evans, A. (1996). Specific involvement of human parietal systems and the amygdala in the perception of biological motion. J. Neurosci. *16*, 3737–3744.

Bortoletto, M., Mattingley, J.B., and Cunnington, R. (2011). Action intentions modulate visual processing during action perception. Neuropsychologia *49*, 2097–2104.

Bradley, M.M., Sabatinelli, D., Lang, P.J., Fitzsimmons, J.R., King, W., and Desai, P. (2003). Activation of the visual cortex in motivated attention. Behav. Neurosci. *117*, 369–380.

Brass, M., Schmitt, R.M., Spengler, S., and Gergely, G. (2007). Investigating action understanding: inferential processes versus action simulation. Curr. Biol. *17*, 2117–2121.

Brown, E.C., and Brüne, M. (2012). The role of prediction in social neuroscience. Front Hum Neurosci *6*, 147.

Bubic, A., von Cramon, D.Y., Jacobsen, T., Schröger, E., and Schubotz, R.I. (2009). Violation of expectation: neural correlates reflect bases of prediction. J. Cogn. Neurosci. *21*, 155–168.

Buckholtz, J.W., Asplund, C.L., Dux, P.E., Zald, D.H., Gore, J.C., Jones, O.D., and Marois, R. (2008). The neural correlates of third-party punishment. Neuron *60*, 930–940.

Calder, A.J., Beaver, J.D., Winston, J.S., Dolan, R.J., Jenkins, R., Eger, E., and Henson, R.N. (2007). Separate coding of different gaze directions in the superior temporal sulcus and inferior parietal lobule. Curr. Biol. *17*, 20–25.

Carter, E.J., Hodgins, J.K., and Rakison, D.H. (2011). Exploring the neural correlates of goal-directed action and intention understanding. Neuroimage *54*, 1634–1642.

Chang, L.J., and Sanfey, A.G. (2013). Great expectations: neural computations underlying the use of social norms in decision-making. Soc. Cogn. Affect. Neurosci. *8*, 277–284.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav. Brain Sci. *36*, 181–204.

Cloutier, J., Gabrieli, J.D.E., O'Young, D., and Ambady, N. (2011). An fMRI study of violations of social expectations: when people are not who we expect them to be. Neuroimage *57*, 583–588.

Cohen, J.D., Perlstein, W.M., Braver, T.S., Nystrom, L.E., Noll, D.C., Jonides, J., and Smith, E.E. (1997). Temporal dynamics of brain activation during a working memory task. Nature *386*, 604–608.

Costantini, M., Galati, G., Ferretti, A., Caulo, M., Tartaro, A., Romani, G.L., and Aglioti, S.M. (2005). Neural systems underlying observation of humanly impossible movements: an FMRI study. Cereb. Cortex *15*, 1761–1767.

Csibra, G., and Gergely, G. (2007). 'Obsessed with goals': functions and mechanisms of teleological interpretation of actions in humans. Acta Psychol. (Amst.) *124*, 60–78.

de Gardelle, V., Waszczuk, M., Egner, T., and Summerfield, C. (2013). Concurrent repetition enhancement and suppression responses in extrastriate visual cortex. Cereb. Cortex *23*, 2235–2244.

de-Wit, L., Machilsen, B., and Putzeys, T. (2010). Predictive coding and the neural response to predictable stimuli. J. Neurosci. *30*, 8702–8703.

Deen, B., and Saxe, R.R. (2012). Neural correlates of social perception: The posterior superior temporal sulcus is modulated by action rationality, but not animacy. Proceedings of the 33rd Annual Cognitive Science Society Conference 276–81.

den Ouden, H.E.M., Friston, K.J., Daw, N.D., McIntosh, A.R., and Stephan, K.E. (2009). A dual role for prediction error in associative learning. Cereb. Cortex *19*, 1175–1185.

den Ouden, H.E.M., Daunizeau, J., Roiser, J., Friston, K.J., and Stephan, K.E. (2010). Striatal prediction error modulates cortical coupling. J. Neurosci. *30*, 3210–3219.

den Ouden, H.E.M., Kok, P.P., and de Lange, F.P.F. (2012). How prediction errors shape perception, attention, and motivation. Front. Psychol. *3*, 548.

Desimone, R.R., Albright, T.D.T., Gross, C.G.C., and Bruce, C.C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. J. Neurosci. *4*, 2051–2062.

DiCarlo, J.J., Zoccolan, D., and Rust, N.C. (2012). How does the brain solve visual object recognition? Neuron *73*, 415–434.

Drugowitsch, J.J., and Pouget, A.A. (2012). Probabilistic vs. non-probabilistic approaches to the neurobiology of perceptual decision-making. Curr. Opin. Neurobiol. *22*, 963–969.

Dunne, S., and O'Doherty, J.P. (2013). Insights from the application of computational neuroimaging to social neuroscience. Curr. Opin. Neurobiol. *23*, 387–392.

Egner, T., Monti, J.M., and Summerfield, C. (2010). Expectation and surprise determine neural population responses in the ventral visual stream. J. Neurosci. *30*, 16601–16608.

Ewbank, M.P., Henson, R.N., Rowe, J.B., Stoyanova, R.S., and Calder, A.J. (2013). Different neural mechanisms within occipitotemporal cortex underlie repetition suppression across same and different-size faces. Cereb. Cortex *23*, 1073–1084.

Fehr, E., and Camerer, C.F. (2007). Social neuroeconomics: the neural circuitry of social preferences. Trends Cogn. Sci. *11*, 419–427.

Feldman, H., and Friston, K.J. (2010). Attention, uncertainty, and free-energy. Front Hum Neurosci *4*, 215.

Felleman, D.J., and Van Essen, D.C. (1991). Distributed hierarchical processing in the primate cerebral cortex. Cereb. Cortex *1*, 1–47.

Fiorillo, C.D., Tobler, P.N., and Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. Science *299*, 1898–1902.

Friston, K. (2005). A theory of cortical responses. Philos. Trans. R. Soc. Lond. B Biol. Sci. *360*, 815–836.

Friston, K.J. (2009). Modalities, modes, and models in functional neuroimaging. Science *326*, 399–403.

Friston, K. (2010). The free-energy principle: a unified brain theory? Nat. Rev. Neurosci. *11*, 127–138.

Frith, C.D., and Frith, U. (2000). The Physiological Basis of Theory of Mind. In Understanding Other Minds: Perspective From Developmental Soc Neurosci, S. Baron-Cohen, H. Tager-Flusberg, and D. Cohen, eds. (Oxford: Oxford University Press), pp. 335–356.

Furl, N., van Rijsbergen, N.J., Treves, A., Friston, K.J., and Dolan, R.J. (2007). Experience-dependent coding of facial expression in superior temporal sulcus. Proc. Natl. Acad. Sci. USA *104*, 13485–13489.

Gergely, G., and Csibra, G. (2003). Teleological reasoning in infancy: the nai;ve theory of rational action. Trends Cogn. Sci. *7*, 287–292.

Graf, M., Reitzner, B., Corves, C., Casile, A., Giese, M., and Prinz, W. (2007). Predicting point-light actions in real-time. Neuroimage *36*(*Suppl 2*), T22–T32.

Gregory, R.L. (1980). Perceptions as hypotheses. Philos. Trans. R. Soc. Lond. B Biol. Sci. *290*, 181–197.

Grill-Spector, K., Henson, R., and Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. Trends Cogn. Sci. *10*, 14–23.

Grossman, E.D., Battelli, L.L., and Pascual-Leone, A.A. (2005). Repetitive TMS over posterior STS disrupts perception of biological motion. Vision Res. *45*, 2847–2853.

Grossman, E.D., Jardine, N.L., and Pyles, J.A. (2010). fMR-adaptation reveals invariant coding of biological motion on the human STS. Front Hum Neurosci *4*, 15.

Hamilton, D.L., and Sherman, S.J. (1996). Perceiving persons and groups. Psychol. Rev. *103*, 336–355.

Hayden, B.Y., Heilbronner, S.R., Pearson, J.M., and Platt, M.L. (2011). Surprise signals in anterior cingulate cortex: neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. J. Neurosci. *31*, 4178–4187.

He, D., Kersten, D., and Fang, F. (2012). Opposite modulation of high- and low-level visual aftereffects by perceptual grouping. Curr. Biol. *22*, 1040–1045.

Herrmann, K., Montaser-Kouhsari, L., Carrasco, M., and Heeger, D.J. (2010). When size matters: attention affects performance by contrast or response gain. Nat. Neurosci. *13*, 1554–1559.

Hesselmann, G.G., Sadaghiani, S.S., Friston, K.J.K., and Kleinschmidt, A.A. (2010). Predictive coding or evidence accumulation? False inference and neuronal fluctuations. PLoS ONE *5*, e9926.

Hillyard, S.A., Vogel, E.K., and Luck, S.J. (1998). Sensory gain control (amplification) as a mechanism of selective attention: electrophysiological and neuroimaging evidence. Philos. Trans. R. Soc. Lond. B Biol. Sci. *353*, 1257–1270.

Ishai, A., Pessoa, L., Bikle, P.C., and Ungerleider, L.G. (2004). Repetition suppression of faces is modulated by emotion. Proc. Natl. Acad. Sci. USA *101*, 9827–9832.

Izuma, K., Saito, D.N., and Sadato, N. (2008). Processing of social and monetary rewards in the human striatum. Neuron *58*, 284–294.

Jastorff, J., Clavagnier, S., Gergely, G., and Orban, G.A. (2011). Neural mechanisms of understanding rational actions: middle temporal gyrus activation by contextual violation. Cereb. Cortex *21*, 318–329.

Jehee, J.F., and Ballard, D.H. (2009). Predictive feedback can account for biphasic responses in the lateral geniculate nucleus. PLoS Comput. Biol. *5*, e1000373.

Jellema, T., and Perrett, D.I. (2003a). Cells in monkey STS responsive to articulated body motions and consequent static posture: a case of implied motion? Neuropsychologia *41*, 1728–1737.

Jellema, T.T., and Perrett, D.I.D. (2003b). Perceptual history influences neural responses to face and body postures. J. Cogn. Neurosci. *15*, 961–971.

Jellema, T., Baker, C.I., Wicker, B., and Perrett, D.I. (2000). Neural representation for the perception of the intentionality of actions. Brain Cogn. *44*, 280–302.

Jones, R.M., Somerville, L.H., Li, J., Ruberry, E.J., Libby, V., Glover, G., Voss, H.U., Ballon, D.J., and Casey, B.J. (2011). Behavioral and neural properties of social reinforcement learning. J. Neurosci. *31*, 13039–13045.

Kable, J.W., and Chatterjee, A. (2006). Specificity of action representations in the lateral occipitotemporal cortex. J. Cogn. Neurosci. *18*, 1498–1517.

Kanwisher, N. (2010). Functional specificity in the human brain: a window into the functional architecture of the mind. Proc. Natl. Acad. Sci. USA *107*, 11163–11170.

Kanwisher, N., McDermott, J., and Chun, M.M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. Neurosci. *17*, 4302–4311.

Keller, G.B., Bonhoeffer, T., and Hübener, M. (2012). Sensorimotor mismatch signals in primary visual cortex of the behaving mouse. Neuron *74*, 809–815.

Kilner, J.M., and Frith, C.D. (2008). Action observation: inferring intentions without mirror neurons. Curr. Biol. *18*, R32–R33.

Kok, P.P., Jehee, J.F.M.J., and de Lange, F.P.F. (2012a). Less is more: expectation sharpens representations in the primary visual cortex. Neuron *75*, 265–270.

Kok, P.P., Rahnev, D.D., Jehee, J.F.M.J., Lau, H.C.H., and de Lange, F.P.F. (2012b). Attention reverses the effect of prediction in silencing sensory signals. Cereb. Cortex *22*, 2197–2206.

Koster-Hale, J., Saxe, R., Dungan, J., and Young, L.L. (2013). Decoding moral judgments from neural representations of intentions. Proc. Natl. Acad. Sci. USA *110*, 5648–5653.

Kourtzi, Z., and Kanwisher, N. (2000). Activation in human MT/MST by static images with implied motion. J. Cogn. Neurosci. *12*, 48–55.

Krueger, J., and Clement, R.W. (1994). The truly false consensus effect: an ineradicable and egocentric bias in social perception. J. Pers. Soc. Psychol. *67*, 596–610.

Lane, R.D., Chua, P.M., and Dolan, R.J. (1999). Common effects of emotional valence, arousal and attention on neural activation during visual processing of pictures. Neuropsychologia *37*, 989–997.

Lin, A., Adolphs, R., and Rangel, A. (2012). Social and monetary reward learning engage overlapping neural substrates. Soc. Cogn. Affect. Neurosci. *7*, 274–281.

Logothetis, N.K., and Sheinberg, D.L. (1996). Visual object recognition. Annu. Rev. Neurosci. *19*, 577–621.

Ma, N., Vandekerckhove, M., Baetens, K., Van Overwalle, F., Seurinck, R., and Fias, W. (2012). Inconsistencies in spontaneous and intentional trait inferences. Soc. Cogn. Affect. Neurosci. *7*, 937–950.

Malle, B.F. (1999). How people explain behavior: a new theoretical framework. Pers. Soc. Psychol. Rev. *3*, 23–48.

Manera, V., Becchio, C., Schouten, B., Bara, B.G., and Verfaillie, K. (2011). Communicative interactions improve visual detection of biological motion. PLoS ONE *6*, e14594.

Mars, R.B., Sallet, J., Neubert, F.-X., and Rushworth, M.F.S. (2013). Connectivity profiles reveal the relationship between brain areas for social cognition in human and monkey temporoparietal cortex. Proc. Natl. Acad. Sci. USA *110*, 10806–10811.

Martinez-Trujillo, J.C.J., and Treue, S.S. (2004). Feature-based attention increases the selectivity of population responses in primate visual cortex. Curr. Biol. *14*, 744–751.

Maunsell, J.H., and Newsome, W.T. (1987). Visual processing in monkey extrastriate cortex. Annu. Rev. Neurosci. *10*, 363–401.

Mende-Siedlecki, P., Cai, Y., and Todorov, A. (2012). The neural dynamics of updating person impressions. Soc. Cogn. Affect. Neurosci. *8*, 623–631.

Meyer, T., and Olson, C.R. (2011). Statistical learning of visual transitions in monkey inferotemporal cortex. Proc. Natl. Acad. Sci. USA *108*, 19401–19406.

Meyer, M., and Sauerland, U. (2009). A pragmatic constraint on ambiguity detection. Nat. Lang. Linguist. Theory *27*, 139–150.

Mitchell, J.P., Macrae, C.N., and Banaji, M.R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. Neuron *50*, 655–663.

Müller, N.G., Strumpf, H., Scholz, M., Baier, B., and Melloni, L. (2013). Repetition suppression versus enhancement—it's quantity that matters. Cereb. Cortex *23*, 315–322.

Mumford, D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. Biol. Cybern. *66*, 241–251.

Murray, S.O., Kersten, D., Olshausen, B.A., Schrater, P., and Woods, D.L. (2002). Shape perception reduces activity in human primary visual cortex. Proc. Natl. Acad. Sci. USA *99*, 15164–15169.

Nakahara, H., Itoh, H., Kawagoe, R., Takikawa, Y., and Hikosaka, O. (2004). Dopamine neurons can represent context-dependent prediction error. Neuron *41*, 269–280.

Neri, P.P., Luu, J.Y.J., and Levi, D.M.D. (2006). Meaningful interactions can enhance visual discrimination of human agents. Nat. Neurosci. *9*, 1186–1192.

Patel, D., Fleming, S.M., and Kilner, J.M. (2012). Inferring subjective states through the observation of actions. Proc. Biol. Sci. *279*, 4853–4860.

Pavlova, M., Staudt, M., Sokolov, A., Birbaumer, N., and Krägeloh-Mann, I. (2003). Perception and production of biological movement in patients with early periventricular brain lesions. Brain *126*, 692–701.

Peelen, M.V.M., and Kastner, S.S. (2011). A neural basis for real-world visual search in human occipitotemporal cortex. Proc. Natl. Acad. Sci. USA *108*, 12125–12130.

Pelphrey, K.A., and Vander Wyk, B.C. (2011). Functional and neural mechanisms for eye gaze processing. In OUP Handbook of Face Perception, A. Calder, G. Rhodes, M. Johnson, and J. Haxby, eds. (Oxford, UK: Oxford University Press), pp. 591–604.

Pelphrey, K.A., Mitchell, T.V.T., McKeown, M.J.M., Goldstein, J.J., Allison, T.T., and McCarthy, G.G. (2003). Brain activity evoked by the perception of human walking: controlling for meaningful coherent motion. J. Neurosci. *23*, 6819–6825.

Pelphrey, K.A., Morris, J.P., and McCarthy, G. (2004). Grasping the intentions of others: the perceived intentionality of an action influences activity in the superior temporal sulcus during social perception. J. Cogn. Neurosci. *16*, 1706–1716.

Perner, J.J., Aichhorn, M.M., Kronbichler, M.M., Staffen, W.W., and Ladurner, G.G. (2006). Thinking of mental and other representations: the roles of left and right temporo-parietal junction. Soc. Neurosci. *1*, 245–258.

Perrett, D.I.D., Smith, P.A.P., Mistlin, A.J.A., Chitty, A.J.A., Head, A.S.A., Potter, D.D.D., Broennimann, R.R., Milner, A.D.A., and Jeeves, M.A.M. (1985). Visual analysis of body movements by neurones in the temporal cortex of the macaque monkey: a preliminary report. Behav. Brain Res. *16*, 153–170.

Poore, J.C., Pfeifer, J.H., Berkman, E.T., Inagaki, T.K., Welborn, B.L., and Lieberman, M.D. (2012). Prediction-error in the context of real social relationships modulates reward system activity. Front Hum Neurosci *6*, 218.

Puce, A.A., Allison, T.T., Bentin, S.S., Gore, J.C.J., and McCarthy, G.G. (1998). Temporal cortex activation in humans viewing eye and mouth movements. J. Neurosci. *18*, 2188–2199.

Rao, R.P., and Ballard, D.H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. *2*, 79–87.

Reynolds, J.H., and Heeger, D.J. (2009). The normalization model of attention. Neuron *61*, 168–185.

Ross, L., Greene, D., and House, P. (1977). The false consensus effect: An egocentric bias in social perception and attribution processes. J. Exp. Soc. Psychol. *13*, 279–301.

Rushworth, M.F., Mars, R.B., and Sallet, J. (2013). Are there specialized circuits for social cognition and are they unique to humans? Curr. Opin. Neurobiol. *23*, 436–442.

Santos, L.R., Flombaum, J.I., and Phillips, W. (2006). The evolution of human mind reading. In Evolutionary Cognitive Neuroscience, S. Platek, ed. (Cambridge: MIT Press), pp. 433–456.

Sapountzis, P., Schluppeck, D., Bowtell, R., and Peirce, J.W. (2010). A comparison of fMRI adaptation and multivariate pattern classification analysis in visual cortex. Neuroimage *49*, 1632–1640.

Saxe, R.R. (2006). Uniquely human social cognition. Curr. Opin. Neurobiol. *16*, 235–239.

Saxe, R., and Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". Neuroimage *19*, 1835–1842.

Saxe, R., and Wexler, A. (2005). Making sense of another mind: the role of the right temporo-parietal junction. Neuropsychologia *43*, 1391–1399.

Saxe, R., Xiao, D.-K., Kovacs, G., Perrett, D.I., and Kanwisher, N. (2004). A region of right posterior superior temporal sulcus responds to observed intentional actions. Neuropsychologia *42*, 1435–1446.

Saygin, A.P., Chaminade, T., Ishiguro, H., Driver, J., and Frith, C. (2012). The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions. Soc. Cogn. Affect. Neurosci. *7*, 413–422.

Schiller, D., Freeman, J.B., Mitchell, J.P., Uleman, J.S., and Phelps, E.A. (2009). A neural mechanism of first impressions. Nat. Neurosci. *12*, 508–514.

Schultz, W. (2010). Dopamine signals for reward value and risk: basic and recent data. Behav. Brain Funct. *6*, 24.

Schultz, W., Dayan, P., and Montague, P.R. (1997). A neural substrate of prediction and reward. Science *275*, 1593–1599.

Senior, C.C., Barnes, J.J., Giampietro, V.V., Simmons, A.A., Bullmore, E.T.E., Brammer, M.M., and David, A.S.A. (2000). The functional neuroanatomy of implicit-motion perception or representational momentum. Curr. Biol. *10*, 16–22.

Senju, A., Johnson, M.H., and Csibra, G. (2006). The development and neural basis of referential gaze perception. Soc. Neurosci. *1*, 220–234.

Seo, H.H., and Lee, D.D. (2012). Neural basis of learning and preference during social decision-making. Curr. Opin. Neurobiol. *22*, 990–995.

Shibata, H., Inui, T., and Ogawa, K. (2011). Understanding interpersonal action coordination: an fMRI study. Exp. Brain Res. *211*, 569–579.

Shultz, S., Lee, S.M., Pelphrey, K., and McCarthy, G. (2011). The posterior superior temporal sulcus is sensitive to the outcome of human and non-human goal-directed actions. Soc. Cogn. Affect. Neurosci. *6*, 602–611.

Spratling, M.W. (2008). Reconciling predictive coding and biased competition models of cortical function. Front Comput Neurosci *2*, 4.

Spratling, M.W. (2010). Predictive coding as a model of response properties in cortical area V1. J. Neurosci. *30*, 3531–3543.

Summerfield, C., Trittschuh, E.H., Monti, J.M., Mesulam, M.M., and Egner, T. (2008). Neural repetition suppression reflects fulfilled perceptual expectations. Nat. Neurosci. *11*, 1004–1006.

Tamir, D.I., and Mitchell, J.P. (2010). Neural correlates of anchoring-and-adjustment during mentalizing. Proc. Natl. Acad. Sci. USA *107*, 10827–10832.

Tenenbaum, J.B., Kemp, C., Griffiths, T.L., and Goodman, N.D. (2011). How to grow a mind: statistics, structure, and abstraction. Science *331*, 1279–1285.

Tobler, P.N., Fiorillo, C.D., and Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. Science *307*, 1642–1645.

Todorovic, A., and de Lange, F.P. (2012). Repetition suppression and expectation suppression are dissociable in time in early auditory evoked fields. J. Neurosci. *32*, 13389–13395.

Todorovic, A., van Ede, F., Maris, E., and de Lange, F.P. (2011). Prior expectation mediates neural adaptation to repeated sounds in the auditory cortex: an MEG study. J. Neurosci. *31*, 9118–9123.

Treue, S., and Martínez Trujillo, J.C. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. Nature *399*, 575–579.

Turk-Browne, N.B., Yi, D.-J., and Chun, M.M. (2006). Linking implicit and explicit memory: common encoding factors and shared representations. Neuron *49*, 917–927.

Wacongne, C., Changeux, J.P., and Dehaene, S. (2012). A neuronal model of predictive coding accounting for the mismatch negativity. J. Neurosci. *32*, 3665–3678.

Weiner, K.S., Sayres, R., Vinberg, J., and Grill-Spector, K. (2010). fMRI-adaptation and category selectivity in human ventral temporal cortex: regional differences across time scales. J. Neurophysiol. *103*, 3349–3365.

Wellman, H.M., Cross, D., and Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. Child Dev. *72*, 655–684.

Wimmer, H., and Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition *13*, 103–128.

Winston, J.S., Henson, R.N., Fine-Goulden, M.R., and Dolan, R.J. (2004). fMRI-adaptation reveals dissociable neural representations of identity and expression in face perception. J. Neurophysiol. *92*, 1830–1839.

Vander Wyk, B.C., Hudac, C.M., Carter, E.J., Sobel, D.M., and Pelphrey, K.A. (2009). Action understanding in the superior temporal sulcus region. Psychol. Sci. *20*, 771–777.

Yamada, M., Camerer, C.F., Fujie, S., Kato, M., Matsuda, T., Takano, H., Ito, H., Suhara, T., and Takahashi, H. (2012). Neural circuits in the brain that are activated when mitigating criminal sentences. Nat Commun *3*, 759.

Young, L.L., and Saxe, R. (2009a). An FMRI investigation of spontaneous mental state inference for moral judgment. J. Cogn. Neurosci. *21*, 1396–1405.

Young, L.L., and Saxe, R. (2009b). Innocent intentions: a correlation between forgiveness for accidental harm and neural activity. Neuropsychologia *47*, 2065–2072.

Young, L.L., Dodell-Feder, D., and Saxe, R. (2010). What gets the attention of the temporo-parietal junction? An fMRI investigation of attention and theory of mind. Neuropsychologia *48*, 2658–2664.

Zaki, J., and Mitchell, J.P. (2011). Equitable decision making is associated with neural markers of intrinsic value. Proc. Natl. Acad. Sci. USA *108*, 19761–19766.

Zhu, L., Mathewson, K.E., and Hsu, M. (2012). Dissociable neural representations of reinforcement and belief prediction errors underlie strategic learning. Proc. Natl. Acad. Sci. USA *109*, 1419–1424.